Spatial Natural Language Generation for Location Description in Photo Captions

Conference or Workshop Item

Spatial Natural Language Generation for Location Description in Photo Captions

Mark M. Hall[1], Christopher B. Jones[2], and Philip Smart[2]

[1] University of Edgehill, United Kingdom
[2] Cardiff University, United Kingdom

Abstract. We present a spatial natural language generation system to create captions that describe the geographical context of geo-referenced photos. An analysis of existing photo captions was used to design templates representing typical caption language patterns, while the results of human subject experiments were used to create field-based spatial models of the applicability of some commonly used spatial prepositions. The language templates are instantiated with geo-data retrieved from the vicinity of the photo locations. A human subject evaluation was used to validate and to improve the spatial language generation procedure, examples of the results of which are presented in the paper.

Keywords: Vague spatial language, Natural language processing, Human-subject experiments, Spatial prepositions, Field-based spatial models, Locative expressions

1 Introduction

Spatial locational expressions (or locative expressions) are used in many aspects of written and spoken communication to describe where things or people are located. Typically the expressions involve the use of spatial relational terms, such as *in* or *at*, to associate the located object to another reference or landmark object. Sometimes an expression may involve the composition of several such spatial relationships, such as "I am on the street in front of the house". In the geographical domain there are many different contexts in which the same spatial relation may be used to refer to phenomena of different types and different scales (buildings, rivers, cities etc). The overloading and contextual dependence of spatial language is well known (e.g. [11], [29]) but presents challenges in developing automated methods to interpret spatial language and to generate spatial language. In this paper we consider the task of generating captions for photos taken with location-aware devices and in that context confine ourselves to urban environments and relatively localised (i.e. small regions of) rural environments. In the locational expressions the located object may be either the photo itself, or the imputed subject of the photo, while the reference objects are confined to named buildings (of various types), streets, settlements and containing regions.

Keeping track of where and when digital photos were taken can be a challenge and has led to interest in methods for automated tagging and captioning (e.g. [16][15]). It is possible to exploit the time and date stamps that most digital cameras provide and, with GPS-enabled devices, it is also possible to access the camera's geographical coordinates. In order to translate the coordinates to a more human-readable form, automated reverse geo-coding services can be used to generate a place name. If other people have taken a photo at a similar location and uploaded it to a photo sharing site such as Flickr then it is possible to suggest tags from the other photos, as proposed by [15]. For previously well photographed scenes a more automatic approach has been presented in which image matching is used to find similar photos and then to inherit the captions of those images [26]. The latter procedure will only work if many people have already taken a photo at the same location. The work presented here differs from these other approaches in automatically generating a complete natural language caption to describe the geographical context of a

photo using locational expressions that are based on analyses of existing captions and on task-specific human-subject experiments. The following is an example of a caption generated fully automatically with our system:

Rijksmuseum photographed at 2.15 pm at the corner of Stadhouderskade and Museumstraat near Spiegelgracht in Amsterdam, Netherlands.

The objective is to emulate typical locational expressions found in image captions in which some care has been taken to describe the geographical context. Evidence of the structure of such captions was obtained from an analysis of the titles of photos uploaded to the Geograph web site. This site is dedicated to providing "geographically representative photographs and information for every square kilometre of Great Britain and Ireland" (www.geograph.org.uk). It is in some contrast to sites such as Flickr in which most photos have only very short captions if any. Some example Geograph captions are listed in Table 1. Our focus here is on the language structure and the selection of spatial prepositions. The appropriate selection of toponyms including reference to their salience [27] is equally important in producing a useful caption and the system presented here uses methods for selecting and ranking toponyms that are described in [30], but these issues are not considered further in this paper as the emphasis is upon caption language structure.

Table 1. Examples of manually generated captions on the Geograph website

Woodland north of Bouverie Avenue, Harnham, Salisbury
The George and Dragon, Castle Street, Salisbury,
Bridge across to Idustrial area at Littlehampton
Cliff Road near Newquay railway station.
Railway bridge over Stratford-on-Avon Canal
Farmland east of Fryern Court Wood
Towards Pendle Hill from York Road, Lanho
Riverside promenade at Brecon
The Monnow above Skenfrith
Postbox on the corner of Linden Road and Gloucester Place
Farmland between Whitsbury & Rockbourne

To understand typical language structures of photo captions and characteristic usage of different prepositions in the context of photo captioning we searched for recurring patterns in language structure, and counted the frequency of occurrence of different spatial prepositions. The pattern analysis led to the creation of a set of language templates of varying levels of complexity. For templates that include a spatial preposition a method is required to determine the most applicable preposition given the configuration of the photo location or subject and the location of candidate toponyms that may serve as referents in a prepositional phrase. Thus decisions need to be made regarding the relative applicability of for example "near <toponym>", "north of <toponym>", "next to <toponym>", "on < street toponym>" and "at the corner of < street toponym A> and < street toponym B>". Knowledge of the applicability of different prepositions was acquired through a set of human subject experiments conducted in a lab and online, in which participants were asked to rate the suitability of a set of prepositions (based on the prior caption analysis) to particular configurations of the located object and a reference location (<toponym>) to which it is related by the preposition. These experiments were similar to those of for example Worboys [32] and Robinson [20] [21]. They differ though in that the subjects were told the context of the task was photo captioning, the scale of map data was adapted to the typical scale found in the caption analysis experiments and the subjects were asked to provide ratings of applicability of given prepositions using values on a Likert scale from 1 to 9. The results of these experiments were used to build, for each preposition, a spatial density field model that fitted a smooth surface to the discrete sample values. The den-

sity field models in combination with prior evidence of the frequency of use of particular prepositions were used to select prepositions to instantiate the language templates.

The main contribution of the work presented here is the design and implementation of a fully automatic natural language photo caption generation procedure that uses a selected set of spatial relations to create a simple description of the geographic context of the photo location. It is based on an analysis of existing caption language patterns to create language templates; analysis of the frequency of use of spatial prepositions in caption language; selection and salience determination of relevant toponyms; modelling of the applicability of a set of spatial prepositions relative to the reference location for the specific context of photo captioning; and instantiation of the language templates with selected spatial prepositions and toponyms.

In the following section we review related work with regard to photo captioning and acquisition and modelling of knowledge of the use of spatial prepositions in geographic contexts. This is followed in section 3 by a description of the caption language analysis while section 4 explains the approach used to create the density field models of prepositions based on human-subject experiments. Section 5 provides an overview of the functionality of the caption generation system that employs the results of the caption analysis and the studies of applicability of spatial prepositions. The section includes a description of the process of selecting and filtering toponyms and an explanation of the selection and instantiation of caption language templates. An initial evaluation of the results of the approach is described in Section 6 along with a discussion of how the results of this evaluation informed various modifications to the final system to take account of insights obtained in the evaluation. The paper concludes in Section 7.

2 Related Work

2.1 Photo captioning

A system with related objectives to our own was the PhotoCompas captioning system [16] which categorises groups of photos according to units of space and time, and can link a photo collection to a neighbouring place using an expression such as "35km S of Los Angeles, CA". While this functionality appears analogous to that provided in our system, it is not clear what prepositions were implemented as only the example preposition of "S" is provided and there is no explanation or discussion of how the spatial prepositional phrase was generated or chosen. Notably their system operates at a relatively coarse scale with the reference locations being cities that may be tens of kilometres distant. This is in contrast to the captions we analysed, in which the reference place (ground location) was usually within 5 km of the photo location and much smaller in dimension than the cities that were used in PhotoCompas.

Another system for organising photo collections was presented in [28] which exploits GPS data to generate times and locations (from a gazetteer) which are related to the photo in terms of distance and cardinal direction. A single structure for annotation is employed and the issues of application of vague spatial language are not addressed. As indicated above several systems, e.g. [15], [26] have been described that attempt to adopt the tags or captions of Flickr photos taken at the same location, but these systems are not concerned with automatic spatial language generation.

2.2 Spatial language

There have been numerous studies concerned with understanding concepts of spatial language and spatial prepositions (e.g. [11, 1]), in particular with regard to the context of use and to distinguishing between frames of reference that may be *relative* to an observer or an object, or based on the properties of an object (*intrinsic*), or *absolute*, such as compass directions, (e.g., [12], [31], [29], [9]). Locative, or locational, expressions are commonly

composed of various forms of figure and ground entities that spatially relate a located objected to a reference object ([29] describes different forms of figure and ground as well as distinguishing between between static and dynamic contexts). While photo captions can contain a wide variety of forms of spatial language we are concerned here primarily with locational expressions that are independent of an observer, i.e. non-deictic, the intention being to generate descriptions from knowledge of the locations of spatial objects and regions in the vicinity of the photo obtained from geo-spatial data sources (if the camera direction is known it would be possible to generate deictic spatial relations but that is not considered here). The focus here is upon external relations (in the sense of for example [12] and [31]) where the photo location or photo subject is the figure (equivalently locatum / located object / trajector), while retrieved places serve as the ground location (relatum / reference location / landmark). We consider ground locations that may be point-based, path-based (roads) or regional. The spatial relations between figure and ground in our system are either independent of a coordinate system, as in linguistic topological relations such as *in*, *next to*, *at the corner of*, and *between* (following the terminology of Levinson [12]), or based on an absolute coordinate system that supports the cardinal directions of *north*, *south*, *east* and *west*. A further aspect of locational descriptions that is relevant to caption generation is that of geographic hierarchies in which a place description will often encompass multiple levels of granularity [19]. This, in combination with evidence from existing photo captions, motivated the inclusion in the present system of support for geographic containment hierarchies when the relevant toponyms are available.

2.3 Modelling the applicability of spatial prepositions

A notable example of creating density field models (or potential fields) of the applicability of spatial prepositions ("regions of acceptability") with evidence from human subject experiments is provided by [13]. Although that work did not have a geographical context, our methods are clearly related in that we also asked subjects to rate the applicability of spatial prepositions at locations relative to a reference location. Of particular relevance to our task of generating locational expressions in specific geographic (as opposed to "table top") contexts are a number of empirical, human-subject studies of the use of vague spatial language concepts that have been concerned with the possibility of fitting models to the experimental data. For example, Robinson conducted studies to acquire fuzzy membership functions to represent the concept of nearness [20] [21] with regard to the relationship between settlements that were mostly tens of kms apart. Using a system that learnt the fuzzy membership function, the subjects were asked to specify the truth or falsehood of nearness for specific instances of pairs of settlements, one of which was the ground location. The latter of these studies emphasised the significant differences between the five participants. Fisher and Orf [5] in a study of the terms *near* and *close* in the context of a university campus found such large variations that they were not able to create a consistent formal model of nearness. Our experiments differed from these latter experiments in using a Likert-like scale to record degrees of applicability of various prepositions and we found that with increasing quantities of data more stable patterns of applicability could be obtained. Gahegan [6] asked subjects to rate the closeness of points to a reference point on a diagram with no absolute scale. In varying the objects in the diagram the study revealed that perceived distance was affected by the presence of neighbouring objects. In our work we do not attempt to consider such "distractor" effects [10], as they are beyond the scope of this study, though it is quite possible that they may have affected the decisions of participants in our human subject studies.

The study in [32] demonstrated the potential value of some alternative approaches to modelling the results of nearness in which human subjects were asked to judge whether it was true that particular landmarks on a university campus were near (or, in other questionnaires, not near) to a specified reference location. The data were modelled in terms

of three-valued logic (corresponding to a broad boundary spatial model), fuzzy distance membership functions and four valued logic. The experiment and analysis was extended in [33] to consider leftness and to adopt Dempster-Shafer belief functions. Using about 22 subjects these studies again revealed individual differences between subjects but also found striking regularities. Our human-subject experiments were analogous to these studies with regard to the type of question asked, though in our experiments judgements were made on a numerical scale from 1 to 9 for multiple spatial relationships, as indicated above. Also, importantly the context of our task was specifically designated as that of photo captioning, unlike any of the previously mentioned studies.

2.4 Spatial natural language processing with density field models

An early example of the application of density field models to generate natural language was presented in Schirra [22] to model the applicability of spatial prepositions to the locations of objects and soccer players in the SOCCER system. It generated a natural language description of a soccer game (based on automatically generated data describing the locations of players and of the ball). Superimposing all fields for a given location, multiple prepositions could be invoked if sufficiently applicable. The form of the field models was based on pre-specified functions that adapt to the shape of referent objects (apparently without any reference to psychometric studies). Our approach is analogous to that of Schirra, in using the models to assess the applicability of different prepositions for the purpose of generating natural language, but we generate the field models from human-subject experiments that were specifically designed for the task of photo captioning.

The use of human subject experiments to build density field models, to represent a variety of spatial preposition terms such as *between*, *near*, *far*, and *to the right of*, was described by Mukerjee et al [14]. Where significant patterns were detected in a particular direction, linear regression methods were used to model the trend of the observations to create ellipsoidal shaped fields. The models were employed to translate from natural language to scene descriptions, and as in Schirra they were superimposed where multiple prepositions applied to the vicinity of a reference object. We have also fitted models to smooth and interpolate our experimental data (with splines and kriging) and we have adopted a similar approach to deciding which is the most applicable of several candidate prepositions (represented by field models).

2.5 Mining text for evidence of the use of spatial language

An example of how natural language texts can be exploited to model the applicability of a prepositional phrase for a specific geographic context was provided by Schockaert et al [23] who analysed text in hotel web sites to produce a fuzzy model of the phrase "within walking distance". Actual distances for instances of the phrase were calculated based on knowledge of a hotel's location in combination with geocoding the location of the named place to which the phrase was applied. An analogous approach was adopted in [7] in which photo captions were mined from the Geograph web site to create spatial field models representing the use of the preposition *near* and the four cardinal directions. Further data on these prepositions was acquired in human subject experiments [8] that employed maps in which participants were asked to rate the applicability of individual prepositions to describe the relationship between multiple named places and a referenced location in a rural setting. Kernel density modelling and kriging [17] interpolation methods, respectively, were used to build the field models for these two studies. The results of [8] are employed in the present paper for captioning in a rural environment.

The study in [8], which was concerned with automated interpretation of the spatial footprints implied by locational expressions in photo captions, revealed three common patterns of caption language. These were a noun phrase giving just a place name, a noun phrase in

combination with a prepositional phrase, such as "Windsor Castle near Eton" and comma separated noun phrases that correspond to a geographical hierarchy such as "Buckingham Palace, London, England". The present study is complementary to that work and builds upon these three basic patterns to generate caption sentences containing locational expressions.

2.6 Natural language generation systems

There has been a considerable body of work on generating natural language and aspects of the overall design of our system build on some well established methods [18] based on which we adopt a data-driven approach to content creation, while discourse planning and linguistic realisation are based on evidence from analysis of existing captions and from human-subject studies. There are many examples of spatial language generation in various domains, notably robotics [24, 10]. In the geographic domain most such systems have focused on navigational instructions, e.g. [3, 4]. While such systems have some generic aspects in common with ours, the approach that we adopt differs with regard to the specific context (which is important given the strongly context-dependent nature of spatial language) and the methods adopted to discover language patterns and to create density models of the applicability of spatial language. An example of language generation that might be regarded as closer in domain to ours is that of a system to create personalised place descriptions for tourists [2], but that work was not concerned with modelling the use of spatial prepositions or the characteristic structure of location descriptions.

3 Analysis of Photo Captions

To gain insight into typical usage of spatial language in photo captions, a set of about 350,000 geocoded photo captions from the Geograph project was analysed. The emphasis of captions in Geograph is upon the geographical context and they are rich in spatial language, in addition to which the photos are usually quite accurately geo-coded. The Geograph dataset was analysed with regard to the frequency of use of different spatial prepositions, the situations in which some prepositions were actually used, with respect to distance and orientation of the photo location relative to a reference location, and the language patterns that were employed in the captions. As with many on-line data sources based on public participation, there is a bias in the data as roughly 90% of the image captions were produced by about 2% of the contributors. To avoid this bias affecting the analyses, only one caption per participant was used in caption language structure analysis. Subsequent equivalent analysis of the full caption set (i.e. with multiple captions from the same contributors) found the same patterns identified with the same relative frequencies, indicating no significant effect of participant bias, which then justified using the full set of images for some of the other quantitative analyses.

3.1 Preposition frequency analysis in Geograph

The analysis of the frequency of use of spatial prepositions found the top ten in descending order to be *at*, *near*, *to*, *on*, *from*, *in*, *north of*, *west of*, *east of* and *south of*. Their numerical frequencies are listed in Table 2 (which includes captions from the same contributors). The numbers were based on analysis of all words in the captions, which were then filtered manually. Of these top ranked prepositions all are used in our system with the exception of *to* and *from* as these tend to be associated either with information about routes that requires additional information sources or with a view-direction specific description, neither of which we attempt to support in this work. We do however support some other spatial relations such as *next to*, *at the corner of* and *between*, where in the latter cases we use retrieved street names to instantiate the prepositional phrases. The knowledge of preposition

frequency is also used in this work to allocate "popularity" weights to prepositions to assist in making decisions about the most appropriate prepositions to employ (in combination with the use of density fields).

Table 2. Most frequent spatial prepositions in Geograph

Preposition	**Frequency**
at	21754
near	18589
to	15476
on	13698
from	12886
in	10754
north	5336
west	5230
east	4763
south	4756

3.2 Figure ground relationships for selected prepositions in Geograph

The preposition *near* and the four cardinal directions were the subject of analysis, in the Geograph dataset, of the distance between photo location and reference location (in the caption) for *near*, and for both distance and angle for cardinal directions relative to a reference location. The measurements of distance and orientation were made following part of speech analysis to detect patterns of the form <subject> <preposition> <toponym> in combination with geocoding the <toponym>, i.e. reference location, and the location of the photo. Details of this form of analysis for the rural use of the cardinal directions were reported in [7], where it was observed that the distances between the photo location and the reference toponym were mostly less than 3km though ranging up to about 5km. The same characteristic distance range has been found in the subsequent analysis of the use of *near* from the same data.

This type of analysis could not be performed automatically with the preposition *at* due to the difficulty of geocoding and disambiguating what were often quite obscure geographic features that were not found in the gazetteers that we employed. A quantitative analysis of *at* was however performed in the human subject experiments described below. The latter experiment was also used to investigate the spatial context of the usage of the path-related prepositions of *on*, *to* and *from*. In the study of Geograph photos it was found that *on* is commonly used in association with a reference toponym that may be visible in the photo, while *to* and *from* often refer to locations that could be quite distant from the visible content of the photo (as for example in the Geograph caption "Gloucester to Swindon Railway, near Minety Cross"). *To* and *from* are therefore harder to employ automatically when attempting to provide the geographic context of a location and are not considered further in the current study.

3.3 Caption language pattern analysis

As a foundation for creating language templates for generating photo captions, the language structure of the Geograph captions was analysed using methods similar to those described in [8], where the aim was automated interpretation of photo captions, as opposed to the generation of photo captions that we are concerned with here. A set of 580 captions, all from different contributors, was derived from the initial collection. To detect language patterns, an iterative process of collocation (bi-gram) analysis was performed on part of speech (POS) tags and subsequently on phrase tags that were substituted for the initially detected tag collocations. The phrase tags were attached to high frequency POS tag collocations that

were identified as English grammatical phrase units. Thus considering the initial POS tag collocation analysis (see Table 3), combinations such as NNP NNP, i.e. two proper nouns, corresponding to two-word place names such as "Chipping Campden", were selected as a noun phrase, designated NPhr, and IN NPhr, i.e. a preposition and a noun phrase, was selected as a prepositional phrase (IPhr) corresponding for example to "near Bussage" or "in Chipping Campden" (see Table 4). Note that NNP IN was not selected as a useful phrase for our purposes as it does not represent a typically meaningful phrase in its own right (it could correspond for example to part of a noun phrase that is a relatively unusual proper name, or the first two words of "Bussage in Goucestershire", or be part of a form of a path description such as "Blogton to Brighton" that we are not seeking to generate in this work).

Table 3. Result of initial collocation analysis of part of speech tags applied to Geograph captions. NNP - proper noun; NN - noun, IN - preposition; DET - determiner ('a' or 'the'); CC - conjunction ('and'); , - comma (',').

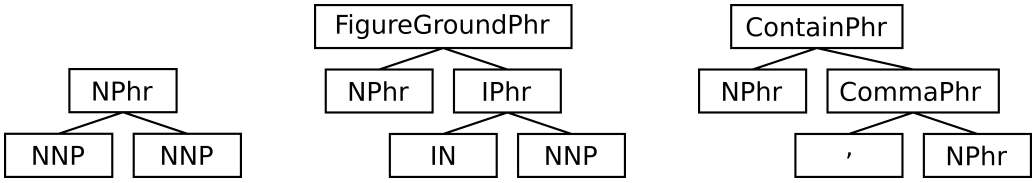
Tag 1	**Tag 2**	**Frequency**
NNP	NNP	632
IN	NNP	149
,	NNP	110
NNP	IN	109
NNP	,	109
NNP	NN	72
DET	NNP	68
NN	IN	62
IN	DET	53
NNP	CC	30

Table 4. Examples of phrase construction generalisation rules, prior to subsequent collocation analysis of the phrase tags. The first seven rules are the result of the first round of generalisation. The subsequent examples illustrate some of the generalisations generated at later rounds. NPhr - noun phrase; IPhr - prepositional phrase; CommaPhr - comma phrase.

Tag 1	**Tag 2**	**Tag 3**	**Generalisation**
NNP	NNP		NPhr
NNP	NN		NPhr
NN	NNP		NPhr
NN	NN		NPhr
DET	NPhr		NPhr
IN	NPhr		IPhr
NPhr	,	NPhr	CommaPhr
NPhr	IPhr		FigureGroundPhr
NPhr	,	CommaPhr	ContainPhr
NPhr	CC	NPhr	ConjunctivePhr

Three iterations of collocation analysis and generalisation phrase creation were performed working from right to left of the sentences (to maintain the structure of noun phrases that may be qualified by a preceding preposition, determiner or adjective). Table 4 illustrates rules generated at the first round (above the separating line) and some of the rules generated at subsequent rounds. The most frequent resulting collocation patterns resulting

from the final generalisation process revealed three major caption patterns that account for 70% of all captions. These are captions that consist of 1) only a noun phrase (NPhr), for example just the toponym "Merthyr Tydfil"; 2) a noun phrase in combination with a prepositional phrase (FigureGroundPhr), for example "Pontsticill Reservoir near Merthyr Tydfil" and 3) captions consisting of a list of comma-separated noun phrases corresponding to a hierarchical toponym such as "Roath Park, Cardiff, Wales" (ContainPhrase). These patterns (Figure 1) provide the basis of a set of building blocks for the caption generation process described in Section 6. It should be noted that while the most common pattern found in Geograph was just the noun phrase representing a single place name, and can therefore be regarded as a typical style of caption, it cannot be regarded as necessarily the most desirable, as it could reflect a minimum effort attitude on the part of the author of the caption. As became apparent in the user evaluation, more complex and hence more informative captions may be preferable to the user. This motivated the creation of captions that combine several templates to create a richer description.



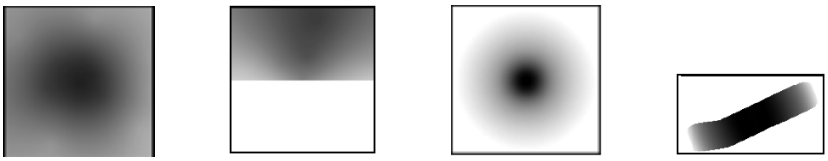
(a) Noun Phrase (b) Figure Ground Phrase (c) Containment Phrase

Fig. 1. The three common caption language patterns detected from analysis of Geograph captions. a) The most common, simple, Noun Phrase structure for a caption, such as 'Merthyr Tydfil'. b) The Figure Ground Phrase consists of two noun phrases linked by a spatial preposition as in 'Pontsticill reservoir near Merthyr Tydfil'. c) The Containment Phrase consists of a set of noun phrases separated by commas as in 'Roath Park, Cardiff, Wales'

4 Field based modelling of spatial prepositions from human subject experiments

In the caption generation system spatial prepositions play a key role, in combination with toponyms, in creating natural-sounding locational expressions. Two human-subject surveys were conducted to acquire knowledge of the applicability of selected spatial prepositions with regard to distances and orientations between figure and ground locations, corresponding respectively to photo locations and reference toponyms. The surveys were based separately on rural and urban contexts. The rural experiment, described previously in [8] for purposes of caption language interpretation, involved 24 undergraduates and university staff. Each participant was shown a map with named point-referenced places and was asked to answer questions of the form of the primer phrase "This photo was taken in < *t*oponym> which is <spatial preposition> Cowbridge", where Cowbridge was centrally located on the map. The overall geographic extent of the map was about 25km which was informed by a prior analysis of the distances between figure and ground locations used in Geograph captioning where it was found that the vast majority of figure and ground locations were within 5km of each other. The context, of taking a photo, was explained but no photos were shown to the participants, as the intention was to describe the geographic content of the camera location, not the content of a photo. For each question the spatial preposition was fixed, while an alphabetically ordered list of toponyms from the map was provided. For each toponym the user was asked to rate on a scale from 1 ("not at all") to 9 ("perfectly") how well it fitted the phrase. The above primer phrase was used to evaluate the prepositions *near* and *north*

of, *south of*, *east of* and *west of*. The result of each answer was a set of figure location points each of which was rated with its applicability for use of the respective preposition relative to the single ground location. For each figure location (corresponding to a named place on the map) the median value of the confidence values was computed, resulting in a set of points with respective confidence values distributed in the vicinity of the single ground location. For each of the prepositions *near* and the four cardinal directions a field model was created using Kriging to perform spatial interpolation between the points. To increase the stability of the resulting fields, given the limited number of figure locations (17), the cardinal direction data points were mirrored across their respective directional axis (east-west and north-south) while the *near* observations were mirrored across both horizontal and vertical axes. Examples of the fields for *near* and *north* are illustrated in Figures 2a and 2b respectively, with the ground location located at the centre of the square (which has truncated the boundaries of the fields in the diagrams).



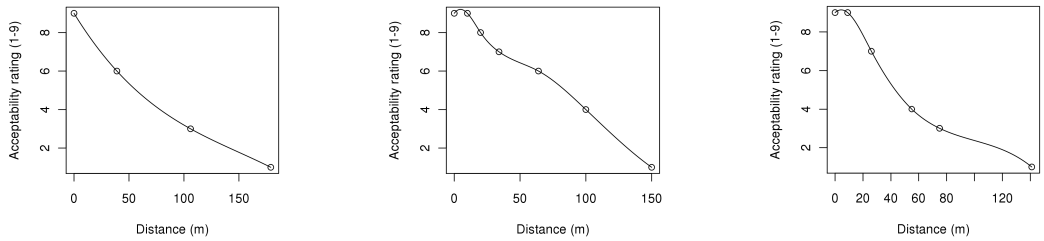
(a) Near (b) North of (c) At the corner (d) Between

Fig. 2. Field models for a) near b) north of, c) at the corner, d) between. The ground location is at the centre of all fields, except d) for which it is a line along the centre path of the strip (corresponding to a road)

The urban experiment (not previously reported) was conducted using a web-based questionnaire that was sent to the same population invited to join the rural experiment. After filtering out participants whose first language was not English, a total of 1042 participants (688 female and 354 male) provided responses.

The usage of the six spatial prepositions *near*, *north of*, *next to*, *at*, *at the corner* and *between* was investigated. The setup for all core questions was the same. On the left side of the screen a square map of a part of the city of Cardiff was displayed. To avoid the participants treating the questionnaire as a map-reasoning task, only a satellite image was displayed. The primer phrase presented to the participants was of the form "Photo taken <spatial expression>", with the spatial expression constructed using one or more toponyms and one of the spatial prepositions, for example "Photo taken near the Wales Millenium Centre". The toponyms in the primer phrase and the photo location were highlighted on the satellite image. The participants were given a nine-point rating scale to rate how applicable the spatial expression was to the spatial configuration shown in the map. The rating elements 1 and 9 were annotated as "does not fit at all" and "fits perfectly" respectively. The core questions were presented in random order, to minimise memory effects. To provide a rough idea of the distances used in the questions, for each spatial preposition the participants would see the closest and most distant photo points first, before seeing the intermediate points. However, to ensure that they would not treat the experiment as a simple geometric reasoning problem the participants were not informed of the order in which the points would be displayed.

Relative to the rural experiments, the urban questionnaire was subject to considerable constraints on the placement of hypothetical photo locations, due to the presence of buildings which were not regarded as available for placing the locations. This resulted in greater sparseness of measurement points (with the exception of the *near* experiment). In the case of the *near* experiment the data values of some individual points differed considerably from the overall pattern for reasons that are unclear but are expected to relate to the higher



(a) At (b) Next To (c) At the corner

Fig. 3. Natural splines fitted through human subject experimental data values obtained for applicability (y-axis; 1 [low] - 9 [high]) of the prepositions 'at', 'next to' and 'at the corner' as a function of distance (x-axis) in the urban setting

density of obstacles and of roads visible in the satellite image, as compared with the relatively simple map employed in the rural experiment. For the cardinal direction *north of* a similar pattern to the rural data was observed, but this was the only cardinal direction that was surveyed in the experiment (in order to control the subject effort of the urban experiment). For these reasons it was decided to employ scaled versions of the rural density fields for these prepositions in their urban context. For the urban *near* data, the closest point to have the lowest median value was about 500m from the ground location, which is one tenth the equivalent distance, i.e. about 5km, that was found with the rural model (though the majority of points for the latter model were within 3km). The scaling factor was therefore chosen as a division by 10 for both *near* and cardinal directions.

In the case of *at*, *next to* and *at the corner* there were too few data points to be subject to Kriging interpolation. Instead a spline function was fitted to the data points. These functions are illustrated for *at*, *next to* and *at the corner* in Figure 3. The function was assumed to be equally applicable in all (radial) directions from a central ground point, producing in the case of the *at the corner* preposition a field of the form illustrated in Figure 2c, in which the ground location is in the centre of the diagram.

For the path like preposition of *between*, the questionnaire used a primer phrase of the form "on street A between street B and street C", with the candidate locations being positioned at various places along the path of street A. The resulting data values were mirrored lengthwise across the centre point of the street to increase the stability of the fitted spline function. To instantiate the vague field for a particular situation the street was assumed to have a width of 20 metres and the spline function was scaled to start and end at the junctions with the streets B and C (wherever they were in practice). The spline function was then used to determine field values for locations on the street according to their distance from the start location, resulting in fields such as that illustrated in Figure 2d.

It is important to stress that the rural and urban human subject experiments were conducted with a view to creating approximate models of the applicability of a working set of prepositions for the specific context of photo captioning. It is well known (as mentioned earlier) that the use of spatial prepositions is highly context dependent and there is no pretence here of having created models that can be regarded as accurate for all types of rural and urban environment with their respective variety of feature types and scales. The experiments were simply part of a pragmatic approach to demonstrating a proof of concept for automated generation of potentially useful locational expressions for photo captions.

5 The Caption Language Generation System

The previous two sections have described the processes of analysing existing photo caption language structure and conducting human subject experiments on the use of various spatial prepositions. The caption analysis resulted in the creation of three forms of caption template, reflecting the three most commonly occurring language patterns, describing subject, relative and containment toponyms. These templates become instantiated with relevant toponyms and spatial prepositions. The human subject studies resulted in the creation of density field models that can be anchored to a toponym location and used to decide the most applicable spatial preposition for the respective photo location. In this section we provide a brief overview of the locational expression generation system that builds on the results of these prior analyses to create the caption. The system applies the process illustrated in Figure 4, which consists of the following four main components:

- The *Meta-Data Extraction* component extracts the location and optional direction information from the image's meta-data.
- The *Meta-Gazetteer* uses the location and direction information to retrieve candidate toponyms that will be used to instantiate the subject, relative and containment language templates. Toponyms are retrieved from a number of sources, ranked based on their salience for captioning, and then filtered as described in section 5.1. The resulting set of toponyms is then passed to the *Captioner*.
- The *Captioner* generates a set of natural language captions using the image's location information, the candidate toponyms, density field models and language templates (sec. 5.2). For the candidate relative toponyms, decisions on instantiating the relative language templates are based on measuring the level of applicability (at the camera location) of all potentially relevant preposition density fields when they are anchored at the toponym location. Templates are merged and, following a linguistic realisation phase, multiple versions of each caption are generated, relating to alternative possible prepositions and toponyms. Each caption is ranked based on a combination of the toponym salience, the applicability of preposition density fields and the popularity of the prepositions.
- The *Meta-Data Embedder* embeds the highest-ranked caption in the image's meta-data.

5.1 Selection and Filtering of Toponyms with a Meta-Gazetteer

The caption language generation system employs a set of caption templates that can be instantiated with spatial prepositions (as explained in the next section) and with toponyms retrieved from the vicinity of the camera location. The toponyms are classified as belonging to one of either subject (S), relative (R) or containment (C) data models. The subject toponyms are ones that occur in a sector in front of the camera, and are only generated if the image's meta-data has orientation direction information (`Dir` choice in Figure 4), the relative toponyms occur anywhere in the circular buffer surrounding the camera location as provided by GPS coordinates, while the containment toponyms provide the geographical regional hierarchy of the photo location. The subject and relative toponyms are allocated salience values that can be used in selection and filtering of names using reverse geocoding and salience measurement methods (Figure 4, *Rank Toponyms*). The toponym reverse geocoding methods employ a meta-gazetteer that accesses multiple data sources, described in [25, 30], while the toponym ranking process uses methods described and evaluated in [30]. In this paper we are concerned primarily with caption language structure rather than the issue of the appropriate selection of toponyms and so in the human subject evaluation experiments described here the toponyms were selected manually for input both to the language generation procedures and to the human-subject map annotators and evaluators.

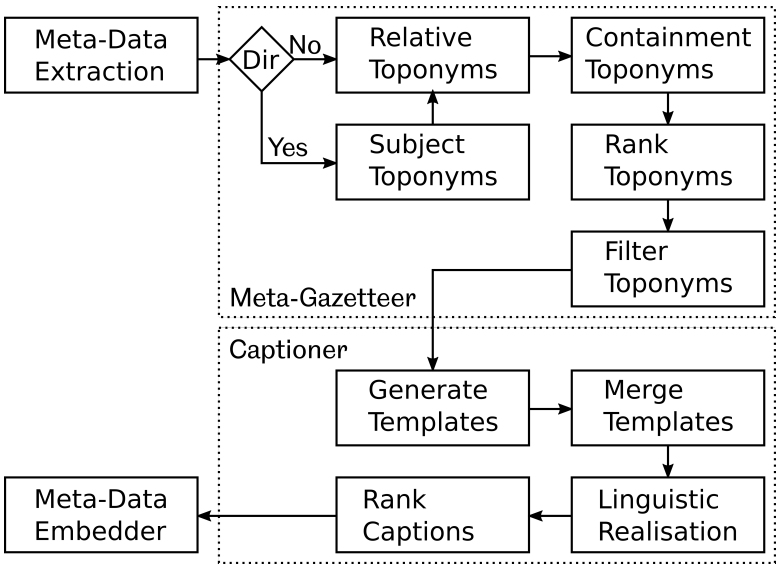


Fig. 4. Overview of components of the caption generation system. The *Meta-Gazetteer* retrieves candidate toponyms that are passed to the *Captioner*, which combines the toponyms with the density field models in order to instantiate the language templates and hence generate the locational expressions. The system follows the same process for each image, except for the *Subject Toponyms* which are only generated if the image's meta-data contains orientation information (`Dir` choice-point).

There were no subject toponyms and all salience values were equal. In the full implemented system all toponyms and their salience values are obtained automatically.

As the data source for subject, relative and containment toponyms could be the same, the possibility arises of the same toponyms occurring in the respective data models. This could result in captions of the form "Cardiff photographed in Cardiff" and thus the toponyms are filtered (Figure 4, *Filter Toponyms*) before being passed to the *Captioner*. First the presence of containment toponyms within the relative and subject toponyms data models is checked. If any are found, they are removed from the S and R data models. Similarly, it is also necessary to remove subject toponyms from the relative toponym models, to avoid captions such as "Wales Millenium Centre near the Wales Millenium Centre".

A further form of filtering is performed to reduce what might be regarded as redundant information, which can occur when a relative toponym provides locational content that is of less semantic salience than the subject toponym. A hypothetical example would be "The Eiffel Tower near the Wagamamma Restaurant", in which the subject is clearly the better known and more unique landmark. This is performed using the semantic salience measures that were generated as part of the captioning system, but are not considered further in this paper as explained above.

5.2 Generating Caption Language Templates

Following creation of the data models containing candidate toponyms to be used in instantiation of the individual caption templates, a discourse modelling phase takes places in which language templates are selected as the basis of what may be multiple candidate caption structures that combine different templates containing different candidate toponyms and candidate spatial prepositions (Figure 4, *Generate Templates*). Which templates are selected depends upon the availability of toponyms and their relative salience. The discourse template models are based on the three major patterns identified in the caption structure analysis (Section 3). Thus the single noun phrase and containment (hierarchy) phrase

patterns form the Containment template, which may consist of one or more toponyms. The figure-ground phrase pattern results in the Relative template which combines one or more toponyms with a spatial preposition. The noun phrase by itself leads to the Subject template, consisting of one or more toponyms that may be combined with an element representing a conjunction phrase. To reflect the common usage of terms such as *on*, *at the corner of* and *between*, referring to path objects as identified in the data mining experiments (and validated in the evaluation experiment described in Section 6) several Road templates are employed (see below). In addition to these, an optional Time template was created with a view to adding additional information within captions, though it was not based on the initial caption structure analysis. The use of the time template is illustrated subsequently in some examples but as it is not a necessary component of the localisation expressions it is not discussed further in this paper.

The templates are combined into the top level discourse model illustrated in Figure 5. According to availability of toponyms and their salience values, the model is populated from left to right.

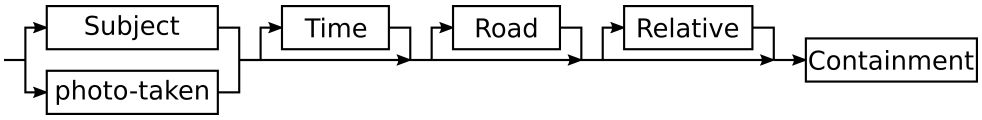


Fig. 5. Top level discourse model. All elements are optional, except the containment element.

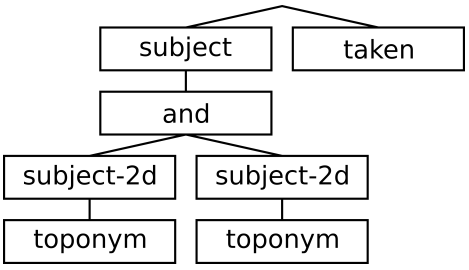


Fig. 6. Template structure for Subject elements.

If subject toponyms are present they are used to instantiate the subject template (Figures 6,7). The 'and' element of the template is used if there is more than one subject toponym. The template includes a 'taken' element that serves as padding to provide more well rounded captions. If a subject is present it can be followed by the word "photographed" (as in the example "Solomon's Temple photographed in the afternoon in Buxton, United Kingdom" - see Table 8 for other examples of captions generated by the automated system), while in the absence of subject toponyms the 'taken' element can still be used and is realised linguistically with the words "Photo taken" (as in the example "Photo taken near Chatsworth House in the Peak District National Park, United Kingdom.").

The road templates implement several phrases that refer to road or street objects. The horizontal support element (Figure 8a), realised by "on <streetname>", can be invoked if the photo location lies within a road as determined by the use of a crisp field model that takes account of road width. All intersections between the road on which the photo is located and other roads are identified and for each intersection that is found an "at the corner" field is instantiated. If the photo location has a field value greater than 0.4 (the cut-off values were determined empirically) then an additional intersection element is

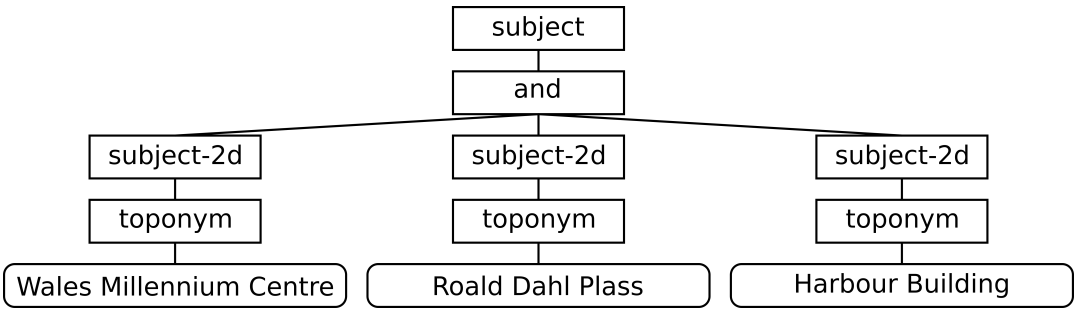
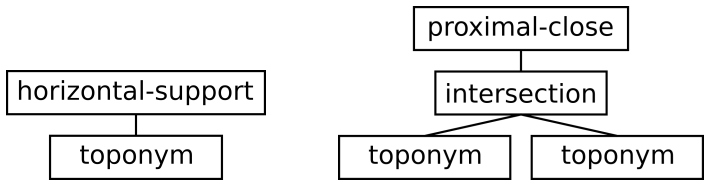


Fig. 7. Template structure for Subject elements with example instantiation. The element would be realised with the phrase "Wales Millennium Centre, Roald Dahl Plass and the Harbour Building".

created and associated with the two streets, resulting in a phrase of the form "at the corner of <streetname> and <streetname>" and wrapped in a proximal close element denoting a short distance from the intersection point (Figure 8b). Finally the procedure generates a "between" template for roads that have at least two intersections and for which the between field has a value exceeding 0.4 at the location of the photo. The full "between" template incorporates a horizontal support element (see Figure 9). It is realised by a phrase of the form "on <streetname> between <streetname> and <streetname>"



(a) *On road* template (b) *At the corner* template

Fig. 8. Road-based templates. a) Basic road-based template realised with a phrase of the form "*on* <streetname>". b *At the corner* template corresponding to the intersection of two named roads.

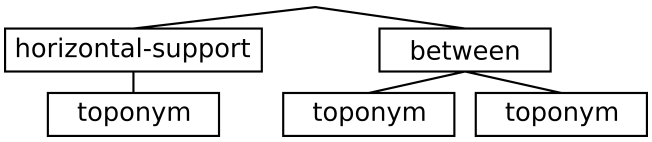


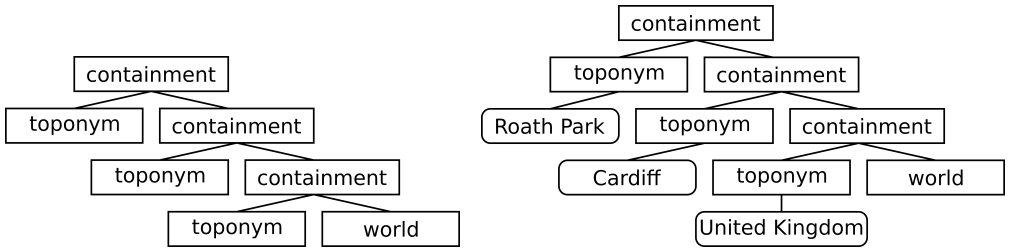
Fig. 9. The *between* road-based template references the names of two roads that intersect the current road.

The relative templates are constructed by iterating over the list of relative toponyms and for each relative toponym instantiating the vague fields for all supported spatial prepositions. A procedure to distinguish between urban and rural situations (using land cover digital map data) is applied to determine the candidate spatial prepositions and hence which field definitions to use. In the rural context, *near* and the cardinal directions are available, while in the urban context *near*, *at*, *next to* and the cardinal directions are instantiated. For each instantiated field (anchored at the candidate toponym location) the value at the photo location is measured and if it is greater than 0.4 a template for the spatial preposition is generated with the toponym as its parameter. In the rural case only fields derived from

Kriging are used whereas in the urban context a distinction is made between the Kriging-based field, for the preposition *near* and the cardinal directions, and spline-based fields for the spatial prepositions *at* and *next to*. Each resulting template is realized by phrases such as "near Little Boddington", "north of Blimpsfield" and "next to St Pauls Cathedral".

An additional filtering step is invoked if at least one road template has been generated, which causes all roads to be ignored when instantiating these vague fields. This avoids generating expressions such as "on Princess street near Princess street".

A single containment template is generated representing the containment hierarchy specified as a list from the most specific to the highest-level containment toponym in the data model. This produces a nested structure as in Figure 10a, in which the containment elements are always interpreted as "the region defined by the left-hand child is contained in the region defined by the right-hand child". The most deeply nested toponym is contained in a special element "world" that is not instantiated to a toponym. This is to ensure that in the structure a containment element always has two child elements, thus simplifying the interpretation, while at the same time providing an explicit end-of-hierarchy marker. The linguistic realisation of the containment template consists in preceding the leftmost toponym (treated as the "root" toponym) with *in* while all subsequent toponyms are concatenated with commas, reflecting the comma phrase pattern that was found in pattern analysis. An example instantiation of the template with three toponyms is illustrated in Figure 10b realised by the containment phrase "in Roath Park, Cardiff, United Kingdom".



(a) *Containment* template (b) Example instantiation of *containment* template

Fig. 10. Containment templates. a) The structure of the containment template. b) Example instantiation of the containment template, which would be realised with the phrase "in Roath Park, Cardiff, United Kingdom"

The final captions are created by merging the templates (Figure 4, *Merge Templates*) and then generating the linguistic realisations of the merged templates (Figure 4, *Linguistic Realisation*). Each caption is given a score based on the sum of individual scores for each subject and relative toponym (Figure 4, *Rank Captions*). In the case of subject toponyms the scores are based only on the salience value attached to the toponym, while for relative toponyms their scores are calculated by multiplying the toponym salience by a factor representing the preposition density field value at the camera location and by a weight associated with each preposition, which is based on its popularity as measured by the data from Geograph on preposition frequency. By default the highest scored caption is selected, but in the automated system the user can be given the option of selecting from other lower ranked captions.

6 Results and evaluation

To evaluate the quality of the natural language expression language caption generation system, the automated captions were compared by human evaluators with human-generated

captions, where the evaluators were not told which one of the captions was computer-generated. The human-generated captions were created for a set of eight locations, four of which were urban, while the other four were rural. For each of the eight locations, three human annotators were given a labelled map and asked to create a caption describing the location of the photo, using the named places on the map and employing a list of spatial prepositions, corresponding to those available to the automated system (note that the choice of prepositions was itself based on the prior study of spatial natural language employed in photo captions, as described in Section 3). In addition to the list of spatial prepositions, the annotators were provided with a list of all toponyms that were on the map. The computer generated captions were created using the same set of toponyms that was available to the human annotators, i.e. all toponyms on their maps.

The same maps used by the caption creators (annotators) were shown to 85 evaluators, with the locations of a notional photo marked on the maps. For each of the human and computer-generated captions (see Tables 5 and 6), which were mixed randomly (i.e. *not* as in these figures), the evaluators were asked to rate on a scale of 1 to 9 how well the caption fitted the given location, where 1 indicated a caption that did not fit at all and 9 represented a perfect caption. The intention was to measure the effectiveness of the caption language in describing the location. For each caption the scores of all evaluators were analysed to calculate the median value and inter-quartile ranges.

It is important to stress that the human generated captions against which the computer generated captions were compared should be regarded as an upper bound. Thus coming close to the manually crafted captions should be regarded as excellent performance.

The first notable outcome of the experiment was that the agreement between evaluators regarding the quality of the human-generated captions was not very high, For almost all human-created captions the inter-quartile range was at least 2 or higher, indicating a large amount of variance in the data.

Table 5. Human-generated and computer generated captions for the urban evaluation experiment. Evaluation areas were: CF 1 & CF 2 – Cardiff; EDI 1 & EDI 2 – Edinburgh. Ann. - Human Annotator generated the caption. Algo. - Algorithm generated the caption.

Area	Source	Caption
CF 1	Ann. 1	On Castle Street, between Cardiff Castle and Cathedral Rd.
	Ann. 2	On Castle St West of Cardiff Castle (or East side of Taff River on Castle St.)
	Ann. 3	On Castle Street near the South West corner of Bute Park
	Algo.	On Castle St near Cardiff Castle in Cardiff
CF 2	Ann. 1	East of Mermaid Quay, South-West of Nat. Assemb. of Wales
	Ann. 2	Between NAW & Mermaid Way
	Ann. 3	In Cardiff Bay between the National Assembly + Mermaid Quay
	Algo.	Near the National Assembly of Wales in Cardiff
EDI 1	Ann. 1	Near Scott Monument, West of Scott Monument
	Ann. 2	Next to Scott Monument
	Ann. 3	On Princess Street between Waverly Bridge and the Royal Scottish Academy
	Algo.	On Princess Street next to the Royal Scottish Academy in Edinburgh
EDI 2	Ann. 1	South of Greyfriars Kirk, on Lauriston Pl
	Ann. 2	On Lauriston Pl near George IV Bridge
	Ann. 3	On Lauriston Pl. West of Univ. of Edinburgh
	Algo.	On Lauriston Pl near the University of Edinburgh in Edinburgh

For the rural urban captions the median values of ratings for the automated captions were 4, 6, 4 and 5 respectively with values ranging between 5 and 8 for the human-generated captions. For the four urban captions the median values for the computer-generated captions

Table 6. Human-generated and computer generated captions for the rural evaluation experiment. Evaluation areas were: BB 1 & BB 2 – Brecon Beacons; PD 1 & PD 2 - Peak District. Ann. - Human Annotator generated the caption. Algo. - Algorithm generated the caption.

Area	Source	Caption
BB 1	Ann. 1	North of Capel Y Ffin
	Ann. 2	Between Velindra and Urishay
	Ann. 3	North of Capel Y Ffin, half-way between Velindra and Urishay
	Algo.	Near Craswall in the Brecon Beacons National Park
BB 2	Ann. 1	North of Coelbren, Near Coelbren
	Ann. 2	Near Coelbren
	Ann. 3	North of Coelbren
	Algo.	Near Coelbren in the Brecon Beacons National Park
PD 1	Ann. 1	Between Wildboardclough and Brandside, North of Quarnford
	Ann. 2	North of Quarnford
	Ann. 3	North of Quarnford half-way between Wildborough and Brandside
	Algo.	Near Dove Head in the Peak District
PD 2	Ann. 1	North of Barbrook Res., West of Totley, Near Owl Bar
	Ann. 2	North of Owl Bar
	Ann. 3	NW of the Owl Bar
	Algo.	Near Owl Bar in the Peak District

were all 5, while they ranged from 6 to 8 for the human-generated captions. The inter quartile ranges for the computer generated captions tended to be high, six of them having a value of 3 and two with a value of 2, demonstrating low agreement between evaluators of the quality of these computer-generated captions.

To provide a qualitative representation of the evaluation, the automated system's ratings were classified into three categories for each caption and each evaluator (see Table 7). The categories are "as good", if the rating for the automatically generated caption is as high or higher than the rating of at least one of the three manually generated captions; "almost as good" if the rating was at most one level lower than the lowest rating of the three manually generated captions and "not as good" if the rating was more than one level below the rating of the lowest rated human generated caption. This was measured for each manual assessment of each photo location tested. The results are illustrated in Figure 12. In summary, we see that the four urban captions were rated "as good" by 35%, 35%, 29%, and 27% of evaluators respectively, with between 45% and 54% of evaluators rating them as either "as good" or "almost as good". For the rural captions there was much more variability in the ratings of the four test cases. In one case the computer generated caption was regarded by 71% of the evaluators as "as good" and by 22% as "almost as good" (thus 93% as either "as good" or "almost as good") and for one of the other captions 61% of evaluators regarded the caption as "as good" and 14% as "almost as good". The other two computer generated captions were rated by 75% and 81% respectively as "not as good" and we discuss the reasons for this shortly.

A further measure of the quality of the automated captions can be found by comparing the spatial prepositional phrases (such as "on Castle Street") and just the selected toponyms of the automated and manually generated captions respectively. With regard to spatial prepositional phrases that combine a preposition and a toponym, in five out of eight of the test locations the automated system generated a prepositional phrase that was the same as that of at least one of the human annotators, examples being "on Princes Street", "on Castle Street" and "near Coelbren" giving a 62.5% success rate for that measure. In six out of eight of the test locations the automated system selected a toponym that was the same as that of at least one human annotator giving a success rate of 75%. In three test

Table 7. Percentages of evaluators' answers in the categories "as good as human" (AG), "almost as good as human" (AAG), "not as good as human" (NAG) for the urban and rural evaluation experiments.

Urban evaluation				Rural evaluation			
Area	**AG**	**AAG**	**NAG**	**Area**	**AG**	**AAG**	**NAG**
CF 1	.35	.19	.46	BB 1	.19	.06	.75
CF 2	.29	.24	.47	BB 2	.71	.22	.07
EDI 1	.35	.16	.49	PD 1	.12	.07	.81
EDI 2	.27	.18	.55	PD 2	.61	.14	.24

cases two toponyms were matched, as for example in the Edinburgh 2 location where both "Princes Street" and "University of Edinburgh" were selected.

Close inspection of the cases where the automated system performed poorly in the evaluation revealed some general limitations that were subsequently rectified. One of these was that the automated system was preferring *near* over *north of* due to a distance decay factor value, based on an analysis of Geograph captions, that was zeroing the cardinal direction fields at a distance notably shorter than the total extent of the near field, so that at the greater distances *near* was always applied rather than a cardinal direction. The Geograph effect for cardinal directions was not in fact seen in human subject experiments and it was decided therefore to omit the scaling factor in the modified version of the system.

Table 8. Examples of captions generated by the fully automated system

Pierhead Building and Norwegian Church photographed in the morning near Wales Millennium Centre in Cardiff, United Kingdom.
Photo taken on Queen Street near the Thistle Parc Hotel in Cardiff City Centre, Cardiff, United Kingdom
Solomon's Temple photographed in the afternoon in Buxton, United Kingdom
Photo taken near Chatsworth House in the Peak District National Park, United Kingdom
Ladybower Reservoir photographed in the early afternoon near Snake Pass in the Peak District National Park, United Kingdom
Rijksmuseum photographed at 2.15pm at the corner of Stadhouderskade and Museumstraat near Spiegelgracht in Amsterdam, Netherlands
Photo taken at the corner of Karolinenstraße and Geyerswörthplatz near Schlenkerla in Bamberg, Germany

Another issue was that in general *near* was being preferred to cardinal directions because the "popularity" weights for each of the cardinal directions reflected their frequency of use in Geograph as described in section 3.1. These weights are independent of the applicability weighting at a given location, which is calculated with a density field. Thus each cardinal direction was used with about a quarter of the frequency of *near* and this had resulted in any cardinal direction being used by the program with only a quarter of the frequency of *near*. This was not really appropriate, as when all four directions are considered, a cardinal direction, i.e. any one of the four directions, should be used with similar frequency to *near* (with each one only occurring about 1 in 4). The popularity weighting for use of cardinal direction was therefore modified to be similar to that of *near*. A similar effect resulted in "between" being used less frequently in the automated system, as it is only applicable in quite specific less common situations, for which the initial low weight was acting as a deterrent to its use in those situations. Its popularity weight was therefore also increased.

The evaluation also indicated that the evaluators preferred more detailed captions, when choosing between all captions including the manually-generated captions. This provides scope for further modifications, whereby multiple cardinal direction phrases could be implemented provided each one exceeded a given threshold of applicability. This could then be expected to emulate a human generated caption such as "north of Capel Y Ffin and east of Velindra". Note however that in the automated system that uses a gazetteer, when many toponyms are available for a given location the captioning system can produce relatively detailed captions as illustrated for example in Table 8.

7 Summary and Conclusions

This paper has presented a set of methods to generate natural language photo captions that employ locational expressions to describe the geographic context of the photo. The captions are based on knowledge of the camera location in combination with access to geo-data resources that relate to the location. There is no reference to the image content. Evidence for the typical structure of caption language was obtained from analysis of systematically authored photo captions on the Geograph web site, resulting in a set of caption language templates corresponding to some of the most common caption patterns that were found. One of the main patterns was a prepositional phrase linking a preposition to one or more toponyms. In order to understand how to select appropriate spatial prepositions that could be applied to particular configurations of distance or orientation between the camera location and named reference locations (toponyms), some human-subject experiments were conducted at rural and urban scales. The results were used to build density field-based spatial models of the applicability of the respective prepositions in the vicinity of the reference toponym. Thus on anchoring the density field to a candidate toponym the applicability of the respective preposition (e.g. west of) could be judged by the value of the field at the camera location.

The resulting locational expression generation system presents a significant step forward as it demonstrates that a purely data-driven approach can successfully be used to model and operationalize spatial natural language in order to automatically generate realistic natural language locational expressions. Compared to existing approaches, the data-driven approach makes it easy to implement a wider range of spatial prepositions, making it possible to create a diversity of natural language locational expressions. This in turn enables the generation of an appropriate expression for a given spatial configuration. The data-driven nature also means that the system can easily be extended to use more spatial prepositions or adapted to other use contexts, by simply providing more quantitative models.

The focus of this paper has been on caption language rather than selection of appropriate toponyms and so an evaluation was performed in which the toponyms were selected manually and provided both to the captioning system and to some human annotators who were asked to create captions using a map containing the selected names. For each notional photo location the automatically generated caption was inserted in a set of human-generated captions and a group of evaluators was asked to judge how well each caption fitted the photo location. Overall just over half the auto-generated captions were judged either "as good" or "almost as good" as the manually generated captions. The best two (of eight) computer generated captions were rated by 93% and 75% of evaluators as either "as good" or "almost as good" as the manually generated captions. This evaluation revealed some problems in the automated system with regard to an inappropriate constraint on the extent of applicability of directional preposition fields and on the expected frequency of use of cardinal directions relative to use of *near*. These problems were corrected in the automated system and the captions re-generated, showing very clear improvement, though no further human-subject evaluation was conducted. Given adequate toponyms for a given location the system can always create well formed captions with good English locational expressions that can include multiple toponyms and multiple prepositions, as illustrated by the examples.

Future work includes conducting further human subject studies on the modified system in a wider range of contexts and using a wider range of spatial prepositions. This will include the use of the automatically generated toponyms that are retrieved with the multi-source "meta-gazetteer". The system presented here has been designed to support realisation of captions in different languages and future work may be conducted on multi-lingual caption generation. An important aspect of the quality of a caption relates to the appropriateness of the selected toponyms, particularly when they might be local landmarks. This issue was not addressed in the paper but has been the subject of related research and further studies will be conducted to determine automatically the salience of particular toponyms, which may be a function of the interests of the owner or user of the photo.

8 Acknowledgements

This work was supported by the EC TRIPOD project (FP6 045335).

References

1. Bateman, J.A., Hois, J., Ross, R.J., Tenbrink, T.: A linguistic ontology of space for natural language processing. Artificial Intelligence 174(14), 1027–1071 (2010)
2. Carolis, B.D., Cozzolongo, G., Pizzutilo, S., Silvestri, V.: Mymap: Generating personalized tourist descriptions. Applied Intelligence 26(2), 111–124 (2007)
3. Dale, R., Geldof, S., Prost, J.: Using natural language generation in automatic route. Journal of Research and practice in Information Technology 36(3), 23 (2004)
4. Dethlefs, N., Y.Wu, Kazerani, A., Winter, S.: Generation of adaptive route descriptions in urban environments. Spatial Cognition & Computation 11(2), 153–177 (2011)
5. Fisher, P.F., Orf, T.M.: An investigation of the meaning of near and close on a university campus. Computers, Environment and Urban Systems 15(1-2), 23–35 (1991)
6. Gahegan, M.: Proximity operators for qualitative spatial reasoning. In: Spatial Information Theory A Theoretical Basis for GIS. pp. 31–44. Springer Berlin / Heidelberg (1995)
7. Hall, M., Jones, C.: Quantifying spatial prepositions: an experimental study. In: Proceedings of the ACM GIS'08. pp. 451–454 (2008)
8. Hall, M., Smart, P., Jones, C.: Interpreting spatial language in image captions. Cognitive Processing 12(1), 67–94 (2011)
9. Herskovits, A.: Semantics and pragmatics of locative expressions. Cognitive Science: A Multidisciplinary Journal 9(3), 341–378 (1985)
10. Kelleher, J., Costello, F.: Applying computational models of spatial prepositions to visually situated dialog. Computational Linguistics 35(2), 271–306 (2009)
11. Landau, B., Jackendoff, R.: "what" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences 16(2), 217–238 (1993)
12. Levinson, S.: Space in language and cognition: Explorations in cognitive diversity. Cambridge: CUP (2003)
13. Logan, G., Sadler, D.: A computational analysis of the apprehension of spatial relations. Language and space pp. 493–529 (1996)
14. Mukerjee, A., Gupta, K., Nautiyal, S., Singh, M., Mishra, N.: Conceptual description of visual scenes from linguistic models. Image and Vision Computing 18(2), 173–187 (2000)
15. Naaman, M., Nair, R.: Zonetag's collaborative tag suggestions: What is this person doing in my phone?. IEEE MultiMedia 15(3), 34–40 (2008)
16. Naaman, M., Song, Y., Paepcke, A., Molina, H.G.: Automatic organization for digital photographs with geographic coordinates. In: JCDL. pp. 53–62 (2004)
17. Oliver, M., Webster, R.: Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information Systems 4(3), 313–332 (1990)
18. Reiter, E., Dale, R.: Building natural language generation systems. Cambridge University Press (2000)
19. Richter, D., Vasardani, M., Stirling, L., Richter, K.F., S. Winter, S.: Zooming in – zooming out: Hierarchies in place descriptions. In: Progress in location-based services, pp. 339–355 (2013)

20. Robinson, V.: Interactive machine acquisition of a fuzzy spatial relation. Computers and Geosciences 16(857–872) (1990)
21. Robinson, V.: Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. Fuzzy Sets and Systems 113(1), 133–145 (2000)
22. Schirra, J.: A contribution to reference semantics of spatial prepositions: The visualization problem and its solution in VITRA. The Semantics of prepositions: from mental processing to natural language processing p. 471 (1993)
23. Schockaert, S., de Cock, M., Kerre, E.: Location approximation for local search services using natural language hints. International Journal of Geographical Information Science 22(3), 315–336 (2008)
24. Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., Brock, D.: Spatial language for human-robot dialogs. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34(2), 154–167 (2004)
25. Smart, P., Jones, C., Twaroch, F.: Multi-source toponym data integration and mediation for a meta-gazetteer service. In: Geographic Information Science (GIScience 2010). pp. 234–248. No. 6292 in Lecture Notes in Computer Science, Springer (2010)
26. Snavely, N., Seitz, S., Szeliski, R.: Modeling the world from internet photo collections. International Journal of Computer Vision 80(2), 189–210 (2007)
27. Sorrows, M., Hirtle, S.: The nature of landmarks for real and electronic spaces. In: Spatial Information Theory., Lecture Notes in Computer Science, vol. 1661, pp. 37–50. (1999)
28. Spinellis, D.: Position-annotated photographs: A geotemporal web. IEEE Pervasive Computing 2(2), 72–79 (2003)
29. Talmy, L.: How language structures space. In: Spatial Orientation, pp. 225–282. New York: Plenum (1983)
30. Tanasescu, V., Smart, P., Jones, C.: Reverse geocoding for photo captioning with a meta-gazetteer. In: SIGSPATIAL 2014. ACM Press (2014)
31. Tenbrink, T.: Reference frames of space and time in language. Journal of Pragmatics 43, 704–722 (2011)
32. Worboys, M.: Nearness relations in environmental space. International Journal of Geographic Information Science 15(7), 633–651 (2001)
33. Worboys, M., Duckham, M., Kulik, L.: Commonsense notions of proximity and direction in environmental space. Spatial Cognition & Computation 4(4), 285–312 (2004)